ABSTRACT

A method and apparatus are provided for removing noise from a first digital representation of data images and noise images of a document, including digitally scanning the document so as to produce the first digital representation of all the images of the document, including the data images and the noise images. After de-skewing the image representation, objects are built from a reduced-resolution representation of the scanned representation. Objects identified as picture objects are included in a mask which is logically ANDed with the de-skewed representation of the scanned document. All objects are added to an object list and initially marked as noise. Objects identified as text objects or geometric objects are marked as data objects. Objects identified as picture objects are included in a mask which is logically ANDed with the de-skewed representation to eliminate all other objects. Objects marked as data objects are added to that representation to provide the de-skewed, de-speckled representation of the scanned document.